图书情报工作
第63卷 第23期 2019年12月
ChinaXiv合作期刊

# 社区画像研究综述*

刘蕾蕾[1,2] 王胜涛[3] 胡正银[1,2]
[1] 中国科学院成都文献情报中心 成都 610041
[2] 中国科学院大学经济与管理学院图书情报与档案管理系 北京 100190
[3] 江南大学糖化学与生物技术教育部重点实验室 无锡 214122

**摘要：**[**目的/意义**]社区画像对于解决社交网络信息过载问题,实现深层次的个性化知识服务意义重大。针对社区画像研究现状,进行客观的分析与评价,以期为社区画像进一步研究与应用提供思路。[**方法/过程**]通过文献调研与分析,从研究内容、方法体系和应用场景3方面对社区画像进行调研、分析与归纳,评述其研究现状,提出未来的重点研究方向。[**结果/结论**]以分析静态用户数据,采用相似性方法画像为主,聚焦于推荐服务、社区发现等传统应用。当前社区画像研究尚处在起步阶段,其数据对象、研究方法与技术手段都有待丰富,社区画像的发展前景与应用空间广阔,需进一步开拓。

**关键词：**社区画像 用户数据 内容画像 传播画像 社区发现 推荐系统 知识服务
**分类号：**G250
**DOI**:10.13266/j.issn.0252-3116.2019.23.014

chinaXiv:202307.00299v1

微信、微博、Academia 以及 ResearchGate 等社交网络飞速发展,越来越多的用户利用社交网络发布或传播信息、分享体验和观点,寻求建议和合作[1-2]。随着社交网络用户群体的不断扩大,社交网络平台中,用户数据分为用户个人数据、社会关系数据、行为数据与用户生成内容(user generated content,UGC)等[3]。利用画像技术对这些数据进行数据建模与知识挖掘,可从中提炼出有价值的信息和知识,实现深层次的个性化知识服务[4]。现有画像研究多集中在单用户画像(individual user profiling),其通过收集与分析用户数据,以标签形式刻画用户特征,挖掘这些特征的潜在价值信息,进而抽象出用户的信息全貌[5-6]。单用户画像在揭示社交网络整体特征方面存在一些不足,如:①从数据层面上看,单用户画像没有充分利用用户社会关系数据,难以全面刻画用户亲近远疏的社会关系[7];②从技术层面上看,单用户画像难以准确过滤 UGC 中的大量噪音数据,导致画像结果常常存在偏差[8];③从应用层面上看,对社区用户群体进行画像更有利于深层次揭示社区特征,支持更广泛的应用[9]。

为了应对这些挑战,群体画像(group profiling)、社区画像(community profiling)等研究相继出现[8]。现有研究并没有对群体画像和社区画像进行严格区分,因此本文统称为社区画像。社区画像是单用户画像的延伸,具有重要研究意义和应用价值。首先,社区画像可以帮助更直观地区分显式社区(explicit communities)与隐式社区(implicit communities),分析用户聚合行为和动机,辅助社区发现[1]。其次,社区画像可以更准确地过滤 UGC 噪音数据,充分利用用户社会关系数据,完善与丰富单用户画像[8]。此外,社区画像还可更全面、精准地支持群体兴趣跟踪[10]、社区知识可视化[11]、社区排名[12]、推荐系统[13]以及网络营销[14]等应用。

本文首先以"community profil*、group profil*、社区画像、群体画像"等为关键词在谷歌学术、百度学术、知网、Elsevier 和 Springer 等搜索引擎和学术数据库中进行检索,得到相关文献 58 篇。然后从研究内容、方法体系和应用场景 3 个方面对社区画像进行综述。最后总结现有社区画像研究的不足以及未来的发展方向。

* 本文系中国科学院"十三五"信息化专项"面向干细胞领域知识发现的科研信息化应用"(项目编号:XXH13506-203)研究成果之一。
作者简介：刘蕾蕾(ORCID:0000-0002-7269-5855),硕士研究生;王胜涛(ORCID:0000-0001-7883-3924),博士研究生;胡正银(ORCID:0000-0002-5699-9891),副研究员,通讯作者,E-mail:huzy@clas.ac.cn。
收稿日期:2019-04-01 修回日期:2019-06-18 本文起止页码:122-130 本文责任编辑:王传清

## 1 社区画像研究内容

### 1.1 社区画像概念

为了应对单用户画像无法满足群体推荐需求的挑战,J. F. Mccarthy 和 T D. Anagnost[15]提出通过社区画像来挖掘小群体的兴趣偏好,来为一组用户提供推荐服务。如 PolyLens 系统主要向 2－4 人的群组提供电影推荐[16]。随着社交网络的兴起、用户生成内容的激增以及数据挖掘等技术的发展,社区画像的对象由小群体发展成为用户规模更大、信息更为丰富、应用场景更为广泛的社区,如对 Twitter、Blog 等大型社交网络中的社区进行画像[8,13]。

社区画像是一个比较新的研究领域,学术界对其定义还没有形成统一的认识。L. Tang[2]、M. Akbari[8]、Z. W. Yu[17]、K. Ashish[18]和石太彬[19]等从画像目的出发,认为社区画像旨在构建社区描述框架,刻画社区群体共同的特征属性与偏好,揭示社区内涵、特性与功能。D. Q. Zhang 等[20]从画像数据出发,指出社区画像本质是成员信息的集合,其应包括成员个人属性、成员偏好数据、社区形成原因以及社区资源等;I. Christensen 等[21]在 D. Q. Zhang 的基础上进一步指出社区由一系列具有相互关系的成员组成,社区画像还应包含成员的社会关系数据。何娟[22]从画像技术出发,指出社区画像旨在挖掘社区特征,综合运用多种数据挖掘方法,分析具有相似特征的用户群体,提炼出每个群体的共同特征,进而针对不同类别的用户群体分别建立有代表性的典型用户画像。A. Salehi[1]和万腾[23]等从画像应用出发,认为社区画像通过挖掘用户群体的使用习惯、访问兴趣等特征,以支持社区排名、兴趣跟踪和社区可视化等社区层面的应用。2017 年,H. Y. Cai 等[24]规范了社区画像的概念,指出社区画像的本质是用户画像信息的融合,从社区内容和社区交互两方面揭示社区特征,并将社区内容定义为内容画像(content profile),社区交互定义为传播画像(diffusion profile)。社区画像的相关概念如表 1 所示。

表 1 社区画像相关概念

| 角度 | 核心内容 | 相关描述 |
| --- | --- | --- |
| 画像目的 | 发现社区特征 | 旨在发现社区的基础属性和代表整个社区的共同特征[8,18－19] |
| | 反映社区偏好 | 融合单用户画像,以反映社区群体的共同偏好[17] |
| | 揭示社区内涵 | 旨在揭示社区内涵、特性与功能,有助于更好的理解社区[2] |
| 画像数据 | 成员信息集合 | 包括成员属性、偏好、社会关系、社区形成原因以及社区资源等数据[20－21] |
| 画像技术 | 社区特征挖掘 | 综合运用分类、聚类、复杂网络分析、机器学习等数据挖掘技术,挖掘具有相似特征的用户群体,提炼群体的共性特征[22] |
| 画像应用 | 社区层面应用 | 挖掘社区特征,抽象出社区群体的使用习惯、访问兴趣等特征信息,以支持社区排名、社区兴趣跟踪和社区可视化等应用[1,23] |
| 综合 | 画像数据/维度/应用 | 用户画像信息的融合,包括内容画像、传播画像两方面,可以支持基于社区理解的传播行为,画像驱动的社区排名和画像驱动的社区可视化等社区层面的应用[24] |

### 1.2 社区画像模型

根据 K. Ashish 等[18]、H. Y. Cai 等[24]、B. Khalid 等[25]的研究,社区画像模型可分成 4 个部分:①收集数据:从各类社交网络或数据平台中获取用户数据。②形成社区:一是基于用户订阅等显性信息,利用分类算法划分显式社区;二是通过分析用户特征等潜在信息,利用社区发现算法生成隐式社区[20]。③社区画像:基于社区用户数据,利用分类、聚类、复杂网络分析、机器学习等数据挖掘技术,结合各类社区画像方法进行社区画像,以揭示社区特征。④画像应用:展示社区画像的应用场景,如群体推荐、寻求合作与辅助决策等。社区画像模型如图 1 所示:

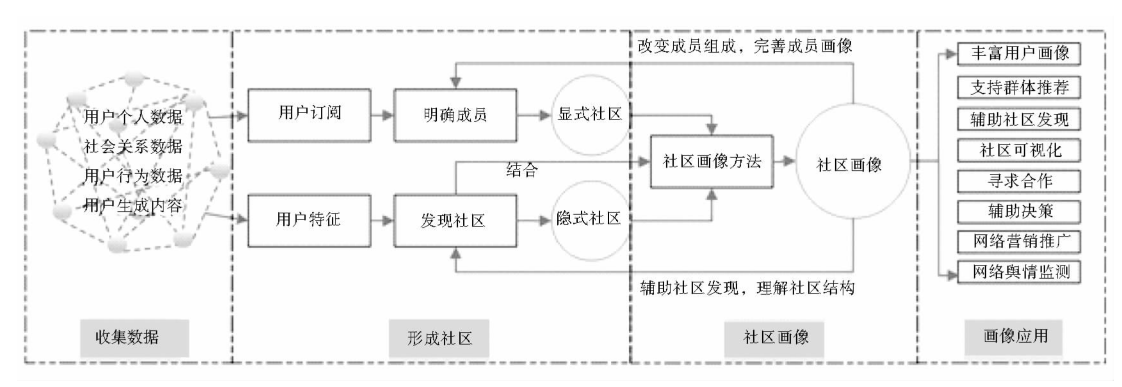



图 1 社区画像模型

chinaXiv:202307.00299v1

ChinaXiv合作期刊

### 1.3 社区画像研究对象

社区画像主要研究对象包括 4 部分(见表 2)。社区画像维度可分为社区内容和社区交互两类;社区内容定义了社区的内涵,如社区的兴趣偏好、行为特征、主题特征等;社区交互则描述社区之间信息传播特征,如社区之间的信息传播模式、社区演化等。社区画像的数据来源主要包括社交网络平台(如微博、Twitter 等)和行业应用平台(如旅游[21]、图书[22]等)。画像技术包括本体、特征表示学习等知识表示技术,特征提取和特征选择等数据降维技术,以及聚类、复杂网络分析、深度学习等数据挖掘技术。社区画像的核心应用场景是推荐服务。此外,社区画像也可应用于网络营销、行为预测、寻求合作和辅助决策等。

表 2　社区画像研究对象

| 画像维度 | 数据集 | 重要画像技术 | 应用场景 |
| --- | --- | --- | --- |
| 社区偏好 | Batch[26] | 语言模型 | 辅助决策;推荐系统 |
| | MovieLens[27] | PDGP | 推荐系统 |
| | Soc/Kids[14] | 分类模型 | 网络营销推广 |
| | Tourism domain[21] | 聚合策略 | 推荐系统 |
| | TV program[17] | 复杂网络分析 | 推荐系统 |
| | 读者数据集[22] | K-means 聚类 | 推荐系统 |
| | 论文数据集[28] | BGLL 算法/作者 - 主题模型 | 辅助决策;寻求合作 |
| 社区行为 | Habrahabr[29] | K-means 聚类 | 丰富用户画像;网络舆情监测 |
| | Twitter[8,13] | 特征表示学习/图聚类/NPLM | 社区可视化;网络营销推广 |
| | 客户消费数据[30] | 云模型聚类 | 网络营销推广 |
| | 通话数据集[19] | NESA 社区发现 | 推荐系统 |
| | 医享网[31] | 概念格聚类 | 推荐系统 |
| 社区主题 | Arxiv Co-Autored[32] | WRS/Chi-Square/BNS/TF-IDF | 辅助社区发现;寻求合作 |
| | Blogcatalog/ Livejounal[2] | 分类模型/特征选择 | 辅助社区发现;寻求合作 |
| | Enron Emails[10] | 机器学习/主题模型 | 辅助决策;寻求合作,推荐系统 |
| | iOS 平台[23] | FCM 聚类 | 网络营销推广 |
| | OJE[33] | PART 机器学习/分类模型 | 辅助社区发现 |
| | Us/UkDataset[1] | GSNMF 聚类 | — |
| | 农业数据集[34] | 本体/FCM 聚类 | 推荐系统 |
| | 音乐数据集[35] | 本体/聚类 | 推荐系统 |
| | 专家系统[25] | — | 寻求合作 |
| 社区传播 | Coauthor/Weibo[36] | CRM/分类算法 | 舆情监测 |
| | Twitter/DBLP[24] | CPD 模型 | 舆情监测;社区可视化 |
| | Weibo[37-39] | COLD/主题模型/图模型聚类 | 舆情监测;网络营销 |

## 2 社区画像方法

从社区形成动机角度出发,依据 H. Tajfel[40]、J. C. Turner 等[41]的社会分类理论,可将社区画像方法分为基于用户相似性画像(user similarities-based profiling,USP)和基于社区差异性画像(community differentiation-based profiling,CDP)两大类。USP 通过分析社区成员共同的兴趣、相近的情感、观点或行为等因素探索社区形成的原因,是研究的热点。USP 又可分为基于单用户画像融合的社区画像与基于用户数据的社区画像两种。基于社区差异性画像方法是通过分析社区内外成员之间的差异,来刻画社区的特征,又可分为基于完整社交网络的差异性画像(differentiation-based group profiling,DGP)和基于社区自身成员的差异性画像(egocentric differentiation - based group profiling,EDGP)两种。社区画像方法见图 2。

### 2.1 基于单用户画像融合的社区画像

基于单用户画像融合的社区画像方法首先基于用户数据形成单用户画像;然后计算不同用户画像间的相似程度;继而将相似的用户画像聚为一类;最后将聚在一起的单用户画像进行融合,生成有代表性典型用户的社区画像[17, 42-43]。该方法的核心在于采取合适的聚合策略来对单用户画像进行融合,最终形成社区画像[17]。聚合策略可分为基于多数的聚合策略、基于共识的聚合策略以及基于边界的聚合策略 3 种[44]。其中,基于多数的策略倾向考虑多数用户的偏好,如相

chinaXiv:202307.00299v1

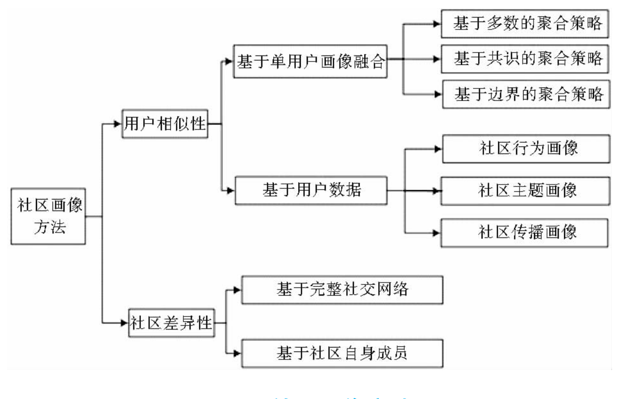


图 2　社区画像方法

对多数投票法、痛苦避免均值策略[15]等;基于边界的策略则只考虑个别用户偏好特征,如最小痛苦策略[16]、最开心策略、最受尊敬策略等;而基于共识的策略则会考虑所有用户的偏好,如平均策略[17],乘法策略[21,45]、波达计数法、科普兰规则、赞同投票、公平策略等。C. Senot 等[42]采用平均策略、最受尊敬策略、多数投票策略、最小痛苦和最开心策略,在一个大型的电视节目数据集上进行实验,发现基于平均策略的聚合效果最好的,而最小痛苦、最开心以及相对多数投票策略几乎是无效的。J. Masthoff[43]对上述 11 种聚合策略进行评价与比较,发现从用户角度出发,用户最喜欢的是平均策略、最小痛苦策略和痛苦避免均值策略,从应用角度出发,乘法策略是表现最好的,其他策略则不能有效反映社区群体偏好特征,显著降低了社区成员的满意度。

该方法适用于社区用户数量较少的场景。然而在大型社交网络中,社区的成员和结构经常变化,此时基于单用户画像融合的社区画像方法则效率较低。此外,该方法未充分利用用户社会关系数据,容易导致画像误差,不能全面揭示社区特征。

## 2.2　基于用户数据的社区画像

基于用户数据的社区画像方法则通过分析社区用户数据,利用相应的画像技术直接生成社区画像[1,24]。该方法充分利用了用户个人数据、行为数据、社会关系数据和 UGC 等各类数据,可以有效提高画像精度,是最常用的社区画像方法。根据画像维度的不同,该方法可分为社区行为画像、社区主题画像和社区传播画像 3 种。

2.2.1　社区行为画像　用户参与社区的行为包括发布信息、分享观点、关注其他用户等,通过分析这些行为数据可揭示社区共同的行为兴趣特征。基于用户的关注、转发与评论行为,M. Akbari 和 T. S. Chua[8]提出了一个包括网络层、交互层、内容层与语义层 4 个层面的社区画像方法。网络层是从用户之间互相关注所形成的关系网络的角度来刻画社区的行为特征,如 A 关注 B,A 和 B 共同关注 C,A 和 B 共同被 C 关注。交互层是分析用户之间的转发和回复等关系数据,如 A 转发 B,A 回复 B 等。内容层是分析用户发布信息的内容之间的包含关系,如 A 发布内容中包含 B 的主题。语义层反应用户共用标签信息,如 A 使用 B 的标签。为了降低 UGC 噪音和数据稀疏性的不利影响,作者还对用户行为特征数据进行降维,综合运用谱聚类算法、相似性约束以及线性回归等形成了社区行为画像。

基于用户评论行为,A. Barysheva 等[29]对博客的讨论参与者进行群体行为画像,综合社区用户特征,如用户数量、写下评论的用户数量,被其他用户评论的数量等,以及博文特征,如用户发表博客的数量、用户评论的博客的数量,用户评论之间的平均距离等,基于 k-means 聚类来发现社区并形成社区行为画像。基于用户浏览行为,万腾[23]从用户粘性和用户活跃度两个方面来提取用户访问行为特征,采用改进的模糊聚类算法来发现社区并形成社区行为画像,通过对用户行为数据的初始隶属度矩阵和样本隶属度权重进行改进,克服了传统模糊聚类算法收敛速度较慢且容易受孤立点影响的局限,取得了较高质量的社区行为画像。

上述画像方法是通过定量分析用户行为数据,生成用户的群体行为特征,进而实现社区行为画像。但用户行为数据存在一定的模糊性和随机性,因此还需要从定性的角度进行补充和完善。姚龙飞和何利力[30]设计一个改进的相似度算法来计算用户定性偏好与定量偏好的相似度,将难以量化的用户行为转化成云模型标签来对社区用户的群体行为进行画像,该方法可有效分析用户的不确定与模糊性行为特征。

2.2.2　社区主题画像　社区主题画像通过生成一系列主题来刻画社区内容特征,其核心是如何准确、高效挖掘社区的主题。主题模型是一系列基于概率模型、旨在发现大规模文档中隐性主题结构方法的统称,于 2003 年由 D. M. Blei[46]提出。在主题模型中,文档可表示为一系列主题的概率分布,而主题则表示成作一系列关键词的概率分布[47]。林燕霞和谢湘生[39]基于隐狄利克雷分布(Latent Dirichlet Allocation,LDA)主题模型挖掘社区用户的兴趣主题,利用余弦相似度和多维标度法(Multidimensional Scaling,MDS)等相似性分析方法对用户及其感兴趣的关键词进行聚类,进而发现群体的主题偏好。孟琳[28]基于作者－主题模型来

chinaXiv:202307.00299v1

ChinaXiv合作期刊

挖掘实验室、科研团队以及科研机构的兴趣主题。J. Marui 等[13]利用神经概率语言模型(Neural Probabilistic Language Model,NPLM)来分析社区内容中广泛存在的"一词多义"现象,发现相同的关键词在不同社区中的不同含义,进而实现更精准的社区主题画像。

除了主题挖掘的算法外,社区主题画像质量还依赖于关键词规范化、结构化与语义化组织程度[48]。A. Salehi 等[1]通过增加关键词词性标注和情感标注,来丰富化关键词的语义信息,进而分析用户深层次的情感信息,细粒度地揭示社区对不同主题事件的情感态度。张海涛等[31]利用概念格分析关键词之间的语义关系,进而挖掘社区主题之间的层级、蕴含等关系,实现多维度的社区主题画像。此外,B. Amini 等[49]、贾伟洋[34]和石季辉等[35]用更复杂的本体来描述社区主题之间的语义关系,实现社区主题之间的自动关联与知识推理。

在大型社交网络中,社区主题的产生、发展、转移和湮灭。动态社区主题画像能及时反映社区主题的发展变化过程,成为研究前沿。L. Tang 等[14]提出一种基于分类的动态社区主题画像方法,该方法采用本体来表示社区中群体主题偏好,利用贪婪算法随着社区内容的变化对本体模型不断修正,实现社区主题的动态更新。

2.2.3 社区传播画像 社区传播画像从信息传播的角度刻画社区特征,根据传播对象的不同可分为用户层面传播画像与社区层面传播画像两种[24,50]。社区传播画像研究的信息通常是主题,通过分析主题在用户或社区层面的传播概率,可以有效地揭示用户及社区的主题偏好,对合作预测、舆情监测等应用具有重要意义。

用户层面信息传播旨在挖掘用户之间的信息传播特征与规律,如 Y. J. Zhu 等[51]通过挖掘用户发表信息的内容及其关联关系,来预测信息在特定用户之间的传播概率;B. D. Wang 等[50]则从用户主题偏好、用户影响力以及主题依赖关系等角度来计算主题在用户之间的传播概率。由于用户行为的不确定性与不稳定性,从单用户层面分析信息传播规律容易出现偏差,因此从社区层面来整体分析信息的传播规律成为研究的热点,如 Y. Han 和 J. Tang[36]通过结合用户主题偏好和社区主题偏好,对信息传播进行建模和预测;Z. T. Hu 等[37]利用动态构建主题的方法,系统分析了文献、交通、音乐、运动以及电影等不同类型社区主题信息的传播路径与传播规律。H. Y. Cai[52]进一步归纳了社区信息传播的三要素,即用户主题偏好、社区主题偏好与主题热度,同时实现了社区主题画像和主题传播画像。

### 2.3 基于社区差异性画像

基于社区差异性画像方法分为 DGP 和 EDGP 两种,其中 DGP 从社交网络整体视角分析与计算社区成员与网络中其他社区成员的差异,而 EDGP 则只考虑社区成员和与其有紧密关系的社区外成员之间的差异。DGP 将社区与社区之外的节点分为两类,通过选取在社区内频繁出现,但在社区外很少出现的特征来进行社区画像[2]。DGP 需要计算整个社交网络的所有特征,其时间复杂度和空间复杂度都很高,效率较低。考虑到社区相对于整个社交网络来说规模较小,L. Tang 等[2]对 DGP 进行了改进,提出了 EDGP 方法。EDGP 并不逐一计算社区与网络中其他社区的差异性,仅将与该社区有紧密联系的部分社区纳入差异性计算范畴。DGP 与 EDGP 方法比较见图 3。L. Tang 等研究发现在社交网络规模较大时,EDGP 和 DGP 社区画像的效果相近,但 EDGP 的效率要高得多。

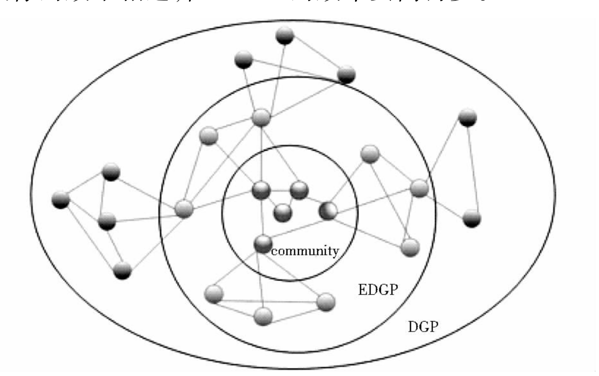



图 3 DGP 与 EDGP 方法比较[2]

### 2.4 社区画像方法比较

目前,基于用户相似性画像是研究与应用的热点,但是该方法倾向于选择普遍流行的特征,其应用存在一定的局限性。如在选择一些宽泛的关键词(如音乐、电影、阅读等)来描述用户或社区偏好时,由于这些关键词在很多社区中都是通用的,此时基于相似性画像方法就不能准确发现社区、刻画社区特征。而基于社区差异性画像方法则倾向于选择使社区区别于其他社区的特征来进行画像,在社区偏好相近或社区特征差异不明显的情况下,表现良好。L. Tang 等[2]在 BlogCatalog 和 LiveJournal 两个大型数据集上比较了这两种方法,发现当特征差异不明显时,基于社区差异性画像方法在特征提取和特征选择上更具备区分度,能更好地发现社区、刻画社区特征。在实际应用中,可根据应用场景和具体需求来选择合适的社区画像方法。社区

chinaXiv:202307.00299v1

刘蕾蕾，王胜涛，胡正银. 社区画像研究综述[J]. 图书情报工作,2019,63(23):122－130.

画像方法比较如表3 所示：

**表3　社区画像方法比较**

| 画像方法 | 画像维度 | 核心算法 | 优点 | 缺点 | 适用场景 | |
|---|---|---|---|---|---|---|
| 基于用户相似性画像 | 基于单用户画像融合 | 社区偏好 | 聚合策略 | 画像算法简单 | 无法准确反映社区群体偏好 | 小规模社交网络 |
| | 基于用户数据 | 社区行为<br>社区主题<br>社区传播 | 分类聚类/主题模型/机器学习等 | 充分利用多类型用户数据，社区画像更精准 | 难以克服用户数据异质性问题 | 用户偏好相似度较高的社交网络 |
| 基于社区差异性画像 | DGP | 社区主题 | 分类/聚类等 | 可用于社区偏好相似度较高的社交网络 | 算法的时间、空间复杂度较高 | 小规模，且社区特征差异不明显的社交网络 |
| | EDGP | 社区主题 | 聚类/复杂网分析等 | 降低 DGP 算法时空复杂度 | 不适用较小规模社区 | 较大规模社交网络 |

## 3　社区画像应用场景

从服务类型的角度来看，社区画像应用场景可分为面向精准推荐服务、面向知识发现服务与面向信息传播服务3 类。

### 3.1　面向精准推荐服务

3.1.1　丰富用户画像　基于单用户画像的个性化推荐服务一方面会因为用户数据稀疏而导致推荐结果不可靠，另一方面由于用户惯性思维容易造成信息茧房，难以满足用户个性化推荐服务需求。而社区画像除了使用用户行为数据外，还综合考虑到社区中典型用户的社会关系、UGC 等数据，可有效规避单用户画像的信息噪音与数据稀疏风险，打破信息茧房，提供更精准、更全面的个性化推荐服务[29,53]。同时社区画像可以包容社交网络中的新用户，为其生成有效的推荐，解决冷启动的问题[12]。如贾伟洋[34]通过构建农业社区画像来为用户进行个性化推荐，不仅有效地提高了推荐准确率，还部分解决了冷启动问题。

3.1.2　支持群体推荐　推荐服务的对象不仅是单个用户，还可以是群体、社区等。社区画像可以有效支持面向群体或社区的精准推荐。许多学者利用社区画像来提高群体推荐服务的满意度，如 E. Ntoutsi 等[54]、C. R. Su 等[55]和 C. Zhang 等[56]利用社区画像进行影视的群体推荐，I. Christensen 等[21]利用社区画像进行旅游领域的群体推荐，何娟[22]利用社区画像进行图书领域的群体推荐。

### 3.2　面向知识发现服务

3.2.1　辅助社区发现　社区发现是社区画像的基础与前提，社区画像则是对社区发现结果的应用与反馈[29,32]。如社区内容画像从社区主题、兴趣偏好和群体行为等不同角度揭示社区内部结构；社区传播画像描述社区之间信息的交互行为，揭示社区外部结构，这些都是社区发现的重要研究内容。此外，社区画像还可以提供更丰富的社区信息，如社区主题、社区偏好等，可以更好地辅助社区发现[24]。如 A. Salehi 等[1]和 H. Y. Cai 等[24]将社区画像与社区发现过程相互结合，利用社区主题画像和情感画像来精准地进行社区发现。

3.2.2　社区信息可视化　社区画像为社区信息可视化提供了更丰富的数据类型与语义信息，支持更直观地揭示社区内容[1,8,57]，可用于帮助分析社区结构、识别重要用户、揭示信息的传播路径等。如 J. D. Cruz 等[11]、H. Y. Cai 等[24]、Z. T. Hu 等[37]在社区画像的基础上，利用社区信息可视化揭示社区之间的交互关系、关联强度以及信息传播途径。

3.2.3　寻求合作与辅助决策　社区画像可帮助用户寻求更有效的合作方式和提供更科学的决策。如 B. Khalid 等[25]提出了一种利用社区画像来提高众包效率的方法，该方法在标记社区用户的专业知识和兴趣偏好基础上，进一步生成群体偏好与知识主题，从而帮助决策者与需求方快速寻找到合作团队，高效解决了众包任务与目标专家不匹配的问题。J. E. A. Gomes 等[32]利用社区画像来分析作者合著网络，提出科学家社区之间的合作模式，并进行合作预测，这在学术合作、知识发现领域有重要的意义。

### 3.3　面向信息传播服务

3.3.1　网络营销推广　社区画像可揭示社区影响力、社区偏好以及社区之间的信息传播模式。这些信息可支持定向广告投放与品牌推广，对网络营销具有重要的应用价值[38,58]。如 Z. T. Hu 等[37]利用社区画像识别具有高影响力的社区，并利用这些社区进行精准网络营销。

3.3.2　网络舆情监测　社区画像可以识别核心用户与热点主题，通过分析他们之间的关系，挖掘社区信息传播模式，进而预测社区用户的信息传播行为，这对舆情监测与管理至关重要。管理层可依据社区用户行

chinaXiv:202307.00299v1

ChinaXiv合作期刊

为、兴趣主题及其影响力等社区画像信息,及时阻断消极的网络舆论,引导积极的社会舆论。如 H. Y. Cai[52]利用社区画像来监测社会事件,实时分析事件发展动态,并预测事件演变。

## 4 结论

社区画像可充分利用用户数据,全面刻画社区特征,为用户提供更精准的推荐服务、深层次知识发现服务与高效的信息传播服务,具有重要研究意义与应用价值。国外对社区画像的研究较早,许多研究将 UGC 与用户关系等多种数据考虑在内,分析的数据类型丰富多样,且注重画像相关基础算法的研究,对社区画像应用研究也较广泛。相比较而言,国内社区画像的研究通常只关注 UGC 或者用户关系等单一维度数据,算法相关研究很少,应用场景主要面向推荐系统。但总体而言,社区画像研究仍处于起步阶段。从数据层面来看,现有研究以分析静态用户数据为主,很少涉及动态社区画像;从技术层面来看,现有研究多采用聚类、主题模型等数据挖掘技术进行小规模的社区画像,而很少利用知识图谱等新型知识技术来进行大规模社区画像;从方法层面来看,现有研究以基于用户相似性画像为主,基于社区差异性画像研究较少;从应用层面来看,目前社区画像研究聚焦于推荐服务、社区发现等传统应用,在以下 4 个方面还有很大的发展空间。

(1)全景式动态社区画像。从数据层面来看,社区画像揭示了社区的结构特征、交互模式、行为模式和发展模式,对未来社区结构预测和演化发展具有重要价值[20]。在大型社交网络中,社区结构、社区成员、社区主题、社区行为以及社区信息传播等社区画像要素都是不断变化的,如何及时、全面反映这些信息对社区画像应用来说非常重要。因此,通过对各类社区数据实时、综合建模,构建全景式动态社区画像,将是未来社区画像研究的热点与难点。

(2)基于知识图谱的社区画像。知识图谱是一种对多源异构数据进行多维度、细粒度知识挖掘与语义关联的新型知识组织技术,是知识互联的基础。从技术层面来看,基于知识图谱的社区画像是一个重要的研究方向,在实践应用中具有重要的意义。基于知识图谱技术进行大规模社区画像,不仅可以充分利用用户数据来挖掘社区的主题网络、传播路径等信息,还可以丰富社区的语义主题,实现社区主题的语义推理与知识发现,为语义搜索、智能问答、推荐系统、数据可视化、大数据分析与决策等应用提供数据支撑。

(3)基于差异性社区画像。随着用户的增加,以及社交网络平台之间合作甚至合并的增多,社交网络规模越来越大,用户或社区之间的特征差异也逐渐变模糊。从方法层面来看,现有画像方法以相似性画像为主,该方法只适用于较小规模且用户差异性明显的社交网络。在大规模社交网络中,用户或社区的特征差异不明显,需要进一步深入研究高效的、差异性敏感的社区画像方法。

(4)社区画像应用场景泛化。社区画像应用前景需进一步泛化。如何通过丰富社区节点语义信息来指导精准社区发现,以及如何将社区画像在推荐服务中的应用进一步泛化,用于支持更加复杂的辅助决策、寻求潜在合作等知识服务,这些都需要结合具体应用需求做进一步探索与尝试。

**参考文献:**

[1] SALEHI A, OZER M, DAVULCU H. Sentiment-driven community profiling and detection on social media[C]// Proceedings of the 29th ACM conference on hypertext and social media. New York: ACM, 2018:229 - 237.

[2] TANG L, WANG X F, LIU H. Group profiling for understanding social structures[J]. ACM transactions on intelligent systems & technology, 2011, 3(1):1 - 25.

[3] 程光曦. SNS 中用户生成内容和行为数据的分析与应用[D]. 北京:北京邮电大学, 2010.

[4] 张钧. 基于用户画像的图书馆知识发现服务研究[J]. 图书与情报,2017(6):60 - 63.

[5] MIDDLETON S E, SHADBOLT N R,DE ROURE D C. Ontological user profiling in recommender systems[J]. ACM transactions on information systems, 2004, 22(1):54 - 88.

[6] ABEL F, GAO Q, HOUBEN G J, et al. Semantic enrichment of twitter posts for user profile construction on the social web[C]//Extended semantic web conference. Berlin: Springer, 2011: 375 - 389.

[7] 关梓骜. 基于大数据技术的用户画像系统的设计与研究[D]. 北京:北京邮电大学,2018.

[8] AKBARI M, CHUA T S. Leveraging behavioral factorization and prior knowledge for community discovery and profiling[C]//Proceedings of the tenth ACM international conference on web search and data mining. New York: ACM, 2017: 71 - 79.

[9] HEIMER C A, HECHTER M. Principles of group solidarity[J]. American Political Science Association, 1989, 83(1):323.

[10] ZHOU W J, JIN H X, LIU Y. Community discovery and profiling with social messages[C]//Proceedings of the 18th ACM SIGKDD international conference on knowledge discovery and data mining. New York: ACM, 2012: 388 - 396.

[11] CRUZ J D, BOTHOREL C, POULET F. Community detection and visualization in social networks: Integrating structural and semantic information[J]. ACM transactions on intelligent systems and tech-

chinaXiv:202307.00099v1

nology (TIST), 2013, 5(1): 1－26.
[12] HAN X, WANG L, FARAHBAKHSH R, et al. CSD: a multi-user similarity metric for community recommendation in online social networks[J]. Expert systems with applications, 2016, 53:14－26.
[13] MARUI J, NORI N, SAKAKI T, et al. Empirical study of conversational community using linguistic expression and profile information[C]//International conference on active media technology. Cham:Springer, 2014: 286－298.
[14] TANG L, LIU H, ZHANG J P, et al. Topic taxonomy adaptation for group profiling[J]. ACM transactions on knowledge discovery from data, 2008, 1(4):1－28.
[15] MCCARTHY J F, ANAGNOST T D. MusicFX: an arbiter of group preferences for computer supported collaborative workouts[C]//Proceedings of the 1998 ACM conference on computer supported cooperative work. New York: ACM, 1998: 363－372.
[16] COSLEY D, KONSTAN J A, RIEDI J. PolyLens: a recommender system for groups of users[C]//ECSCW 2001. Proceedings of the Seventh European conference on computer supported cooperative work. Dordrecht: Springer, 2001: 199－218.
[17] YU Z W, ZHOU X S, HAO Y B, et al. TV program recommendation for multiple viewers based on user profile merging[J]. User modeling and user-adapted interaction,2006, 16(1):63－82.
[18] ASHISH K, JAISWAL U C, UPADHYAY P. Intelligent system using the concept of group profiling by user profiling[J]. International journal of current engineering and technology, 2014, 4(5): 3314－3317.
[19] 石太彬. 基于用户通话记录的社区发现算法与社区画像研究[D]. 杭州:浙江大学,2017.
[20] ZHANG D Q, YU Z Y, GUO B, et al. Exploiting personal and community context in mobile social networks[M]. New York: Springer, 2014: 109－138.
[21] CHRISTENSEN I, SCHIAFFINO S, ARMENTANO M. Social group recommendation in the tourism domain[J]. Journal of intelligent information systems,2016, 47(2):209－231.
[22] 何娟. 基于用户个人及群体画像相结合的图书个性化推荐应用研究[J]. 情报理论与实践,2019,42(1):129－133,160.
[23] 万腾. 基于 iOS 的性能监控和用户操作行为分析研究[D]. 广州:华南理工大学,2017.
[24] CAI H Y, ZHENG V W, ZHU F W, et al. From community detection to community profiling[J]. Proceedings of the Vldb Endowment, 2017, 10(7):817－828.
[25] KHALID B, MARCO B, DANIELA G, et al. Community profiling for crowdsourcing queries[EB/OL]. [2018－12－08]. https://basepub. dauphine. fr/bitstream/handle/123456789/16699/CommunityProfiling_belhajjame_brambilla. pdf? sequence＝2.
[26] DEHGHANI M, AZARBONYAD H, KAMPS J, et al. Generalized group profiling for content customization[C]//Proceedings of the 2016 ACM conference on human information interaction and retrieval. New York: ACM, 2016: 245－248.
[27] TAHA K, ELMASRI R. Personalization with dynamic group profile [C]//Advances in social networks analysis and mining. Turkey: IEEE, 2012: 488－492.
[28] 孟琳. 多源信息融合的机构画像的方法研究[D]. 北京:北京邮电大学,2018.
[29] BARYSHEVA A, PETROV M, YAVORSKIY R. Building profiles of blog users based on comment graph analysis: the Habrahabr. ru case[C]//International conference on analysis of images, social networks and texts. Cham: Springer, 2015: 257－262.
[30] 姚龙飞,何利力. 基于云模型理论的群体用户画像模型[J]. 计算机系统应用,2018, 27(6):53－59.
[31] 张海涛,崔阳,王丹,等. 基于概念格的在线健康社区用户画像研究[J]. 情报学报,2018,37(9):912－922.
[32] GOMES J E A, PRUDÊNCIO R B C, NASCIMENTO A. A comparative study of group profiling techniques in co-authorship networks[C]//Proceedings of 2016 5th Brazilian conference on intelligent systems. Brazil: IEEE, 2016: 373－378.
[33] GOMES J E A, PRUDÊNCIO R B C. Educational social network group profiling: an analysis of differentiation-based methods[EB/OL]. [2019－03－08]. http://www. lbd. dcc. ufmg. br/colecoes/brasnam/2015/011. pdf.
[34] 贾伟洋. 基于群体用户画像的农业信息化推荐算法研究[D]. 咸阳:西北农林科技大学,2017.
[35] 石季辉,于长锐,刘兰娟. 基于领域本体的社区用户兴趣模型[J]. 情报科学, 2011(4):609－613.
[36] HAN Y, TANG J. Probabilistic community and role model for social networks[C]//Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining. New York: ACM, 2015: 407－416.
[37] HU Z T, YAO J J, CUI B, et al. Community level diffusion extraction [C]//Proceedings of the 2015 ACM SIGMOD international conference on management of data. New York: ACM, 2015: 1555－1569.
[38] 何跃,邓姝颖,马玉凤,等. 突发事件中微博用户社群舆情传播特征研究[J]. 情报科学, 2016,34(6):14－18.
[39] 林燕霞,谢湘生. 基于社会认同理论的微博群体用户画像[J]. 情报理论与实践, 2018,41(3):142－148.
[40] TAJFEL H. Social identity and intergroup relations[M]. Cambridge: Cambridge University Press,2010: 22－28.
[41] TURNER J C, HOGG M A, OAKES P J, et al. Rediscovering the social group: a self-categorization theory[J]. British journal of social psychology, 2011, 26(4):347－348.
[42] SENOT C, KOSTADINOV D, BOUZID M, et al. Evaluation of group profiling strategies[C]// Twenty-second international joint conference on artificial intelligence. Spain: DBLP, 2011: 2728－2733.
[43] MASTHOFF J. Group recommender systems: combining individual models[M]. Boston: Springer, 2011: 677－702.
[44] BERNIER C, BRUN A, AGHASARYAN A, et al. Topology of communities for the collaborative recommendations to groups[C]// 3rd international conference on information systems and economic

chinaXiv:202307.00299v1

ChinaXiv合作期刊

intelligence. Tunisia: SIIE, 2010:1 -6.

[45] MASTHOFF J. Group modeling: selecting a sequence of television items to suit a group of viewers[J]. User modeling and user-adapted interaction, 2004, 14(1):37 -85.

[46] BLEI D M, NG A Y, JORDAN M I. Latent dirichlet allocation[J]. Journal of machine learning research, 2003, 3(1): 993 -1022.

[47] 胡正银,方曙,文奕,等.面向TRIZ的专利自动分类研究[J].现代图书情报技术,2015(1):66 -74.

[48] ZHANG Y, PORTER A L, HU Z Y, et al. "Term clumping" for technical intelligence: a case study on dye-sensitized solar cells[J]. Technological forecasting & social change, 2014, 85:26 -39.

[49] AMINI B, IBRAHIM R, OTHMAN M S, et al. A multi-reference ontology for profiling scholars' background knowledge [M]. Cham: Springer, 2014: 35 -46.

[50] WANG B D, WANG C, BU J J, et al. Whom to mention: expand the diffusion of tweets by@ recommendation on micro-blogging systems[C]//Proceedings of the 22nd international conference on World Wide Web. New York: ACM, 2013: 1331 -1340.

[51] ZHU Y J, YAN X R, GETOOR L, et al. Scalable text and link analysis with mixed-topic link models[C]//Proceedings of the 19th ACM SIGKDD international conference on knowledge discovery and data mining. New York: ACM, 2013: 473 -481.

[52] CAI H Y. Making sense of social events by event monitoring, visualization and underlying community profiling[D]. Queensland: The University of Queensland, 2016.

[53] XIE H R, LI Q, MAO X D, et al. Community-aware user profile enrichment in folksonomy[J]. Neural networks, 2014, 58(5): 111 -121.

[54] NTOUTSI E, STEFANIDIS K, NØRVÅG K, et al. Fast group recommendations by applying user clustering[C]//International conference on conceptual modeling. Berlin: Springer, 2012: 126 -140.

[55] SU C R, LI Y W, ZHANG R Z, et al. An adaptive video program recommender based on group user profiles[M]. Berlin: Springer, 2013: 499 -509.

[56] ZHANG C, ZHOU J, XIE W F. A users clustering algorithm for group recommendation[C]//2016 4th international conference on applied computing and information technology/3rd international conference on computational science/intelligence and applied informatics/1st international conference on big data, cloud computing, data science & engineering (ACIT-CSII-BCD). Las Vegas: IEEE, 2016: 352 -356.

[57] LIN Y R, SUN J, CASTRO P, et al. Metafac: community discovery via relational hypergraph factorization[C]//Proceedings of the 15th ACM SIGKDD international conference on knowledge discovery and data mining. New York: ACM, 2009: 527 -536.

[58] WANG Z, ZHANG D Q, ZHOU X S, et al. Discovering and profiling overlapping communities in location-based social networks[J]. IEEE transactions on systems, man, and cybernetics: systems, 2014, 44(4): 499 -509.

**作者贡献说明:**
刘蕾蕾:调查资料,撰写论文;
王胜涛:调查资料,修改论文;
胡正银:指导、修改并审定论文。

## A Literature Review on Community Profiling

Liu Leilei[1,2] Wang Shengtao[3] Hu Zhengyin[1,2]
[1] Chengdu Documentation and Information Center, Chinese Academy of Sciences, Chengdu 610041
[2] Department of Library, Information and Archives Management, School of Economics and Management, University of Chinese Academy of Sciences, Beijing 100190
[3] Key Laboratory of Carbohydrate Chemistry&Biotechnology, Ministry of Education, Jiangnan University, Wuxi 214122

**Abstract:** [**Purpose/significance**] Community profiling is important for solving the overload of social network information and helping to achieve personalized and deep knowledge services. This literature review presents the research status in community profiling, and analyzes the corresponding techniques, methods and applications, and aims to provide ideas for further research and application of community profiling. [**Method/process**] Based on the literature investigation, this paper reviews community profiling from three aspects: research content, techniques and methods, and application scenarios. Moreover, the key features and weaknesses of the discussed techniques and methods are presented and several key research fields for future research are highlighted. [**Result/conclusion**] It is found that the present research focuses on static user data, user similarity methods for profiling, and traditional applications such as recommended services and community discovery. At present, the research on community profiling is still in its infancy, and the data, techniques and methods need to be enriched. It should have good prospects and wide application in the future.

**Keywords:** community profiling user data content profile diffusion profile community detection recommender system knowledge service

chinaXiv:202307.00299v1